

Using *knitr* and \LaTeX for literate lab notes

Boris Veytsman

TUG'2022

A cautionary story

How a student was frustrated trying to repeat research...

Am I sure I can understand my own work in ten years?

How do our colleagues in biomedical and experimental research solve this problem, when billions depend on it?

Laboratory notebooks

A Lab Notebook Is...

- ▶ *Complete record of procedures, reagents, data, and thoughts to pass on to other researchers*
- ▶ *Explanation of why experiments were initiated, how they were performed, and the results*
- ▶ *Legal document to prove patents and defend your data against accusations of fraud*

Philip Ryan (2012). Keeping a Lab Notebook. National Institutes of Health, Office of Intramural Training and Education. URL: [https://www.training.nih.gov/assets/Lab_Notebook_508_\(new\).pdf](https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf)

A classic example: Linus Pauling's notebooks

72 years of work (from 1922 to 1994): <http://scarc.library.oregonstate.edu/coll/pauling/rnb/index.html>

x-ray data made Oct 1922 to June 1923


$K_2Ni_2(SO_4)_3$ $a_0 = 9.078 \text{ \AA}$ $\gamma K, \gamma Ni, \gamma S$
 $NiCl_2$ $a_0 = 3.58 \text{ \AA}$ $0K^+$
 $CaHg_2Br_4$ $a_0 = 11.14$ $Z=32$
 MoO_3
 $(NH_4)_2FeF_6$
 $Ni_2SO_4 \cdot 4H_2O$
 $Ca(NH_4)_6I_3$
 $(NH_4)_2B_2F_6$
 Zn
 $Mg_2Si \cdot Mg_2O_2 \cdot CaHg_2$

Books - Chem papers
 from Oct 15 to Dec 31 1922:
 10 Oct. L.H. with J.D.
 20 Oct. Mg₂Si made NaCl₂
 23 Oct. Cd by statistics
 21 Oct. Cd (NH₄)₂
 Na₂CO₃ · 10H₂O
 K₂Mg(SO₄)₂
 19 Oct. 65C
 23 - Cr
 24 Oct. CoS · 6H₂O · 4H₂O water
 28 - K₂Ni₂(SO₄)₃ made calc
 31 - Ni
 1 Oct. Ni made
 11 Oct. Ni made
 14 Oct. Ta₂O₅ made, white
 15 Oct. K₂HgBr₄ - statistics
 " CaHg₂Br₄ - " " " " " " " "
 10 Dec. MoO₃ made - statistical
 14 Handwritten made
 in scattered 6 weeks

White body $a_0 = 1.1710 \text{ \AA}$ $c = 1.581 \text{ \AA}$ $\gamma K, \gamma Ni, \gamma S$

I got back to work on this last night I decided to extend it to which kind of energy explains properties of metals, assuming square lattice of ions, with volume determined well with volume always proportional to A .

Will do small experiment.



$$V = abc$$

$$E = \frac{a^2}{2ma^2} + \frac{b^2}{2mb^2} + \frac{c^2}{2mc^2} \quad a=b=c$$

$$E = \frac{a^2}{2m} + \frac{b^2}{2m} + \frac{c^2}{2m}$$

Consider $\frac{a^2}{2m} = \frac{b^2}{2m} = \frac{c^2}{2m}$

$$E = \frac{1}{2m} \left(\frac{a^2}{2} + \frac{b^2}{2} + \frac{c^2}{2} \right)$$

Let $abc = 2$ $a=b=c = 2^{1/3}$ $a^2 = 2^{2/3} \approx 1.5874$

$$2mE = \frac{9}{2^3} = 5.1696$$

Let $2mE = 2 + 2 + \frac{9}{2} = 5.2500$

Now let $a^2 = b^2 = c^2 = \frac{9}{2} = 4.5$ $mE = 6.75$

$$a^2 = b^2 = 1.1696 \quad c^2 = 2.9240$$

$$2mE = 8.1299 \approx 7.8 = 5.1300$$

$a = b = 1.0715 \quad c = 1.7110 \quad c/a = 1.581$ Minimum

15 Aug 1923 Amalithia 116

Case on p. 105 - my figure 3.

$$-AS(12 - 20.16 \text{ kcal}) = 27 + 7.06 \log \frac{100}{100}$$

For $\frac{100}{510} = 1$ where $\frac{100}{510} = \frac{1}{5.1}$ $\log = -1.1315$

Now $-AS(12 - 100) = 10.00$

$$-AS(12 - x) = 10.00 - 19.2 = 5.20$$

But this value is quite arbitrary - it depends on the standard state for the gas. Similarly, Wark's argument (p. 6) is faulty.

Lab notes as literate science

Knuth's insight: Your code is for computer. Your prose is for humans. \Rightarrow Literate programming¹.

Research situation: A paper (preprint, presentation) is just an *advertisement* of the research, but not the research. Research is a reproducible *environment* which includes computation and publication². \Rightarrow Lab notes as literate science

¹Donald E. Knuth (1992). *Literate Programming*. CSLI Lecture Notes 27. California: Stanford.

²Jill P. Mesirov (2010). "Accessible Reproducible Research". In: *Science* 327.5964, pp. 415–416. ISSN: 0036-8075. DOI: 10.1126/science.1179653. URL: <http://science.sciencemag.org/content/327/5964/415>.

How do we keep lab notes?

The classic way: bunches of physical notebooks

- ▶ Very versatile: you can put there anything! *But*
- ▶ You cannot search efficiently (where is my `grep`?)
- ▶ Too many dead trees.
- ▶ Not too easy to keep after a couple of decades.

The modern way: electronic records

- ▶ Can be indexed, searched, compact! *But*
- ▶ Can we make them as versatile as physical ones?
- ▶ Can we make writing them as fast as scribbling?

What is in my lab notes? (1)

- ▶ Thoughts and ideas:

It seems that cell diffusion inside a tissue is quite different if a different matrix around the tissue was used. This fact is quite inexplicable from the conventional picture of diffusion borrowed from the molecular physics. Indeed, how would a molecule inside a vessel “know” what is the vessel made of? One expects the measured diffusion not depend on the walls around the molecules.

- ▶ Equations

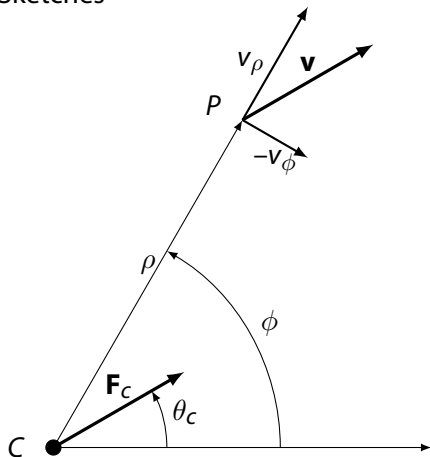
$$a = -\frac{l(l+1)c}{r},$$

$$b = -\frac{dc}{dr} - \frac{c}{r},$$

$$c = -\frac{a}{r} + \frac{db}{dr} + \frac{b}{r}.$$

What is in my lab notes? (2)

► Sketches



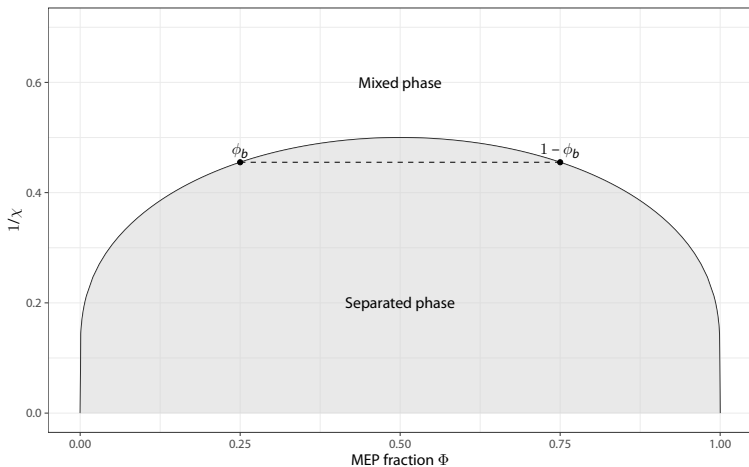
What is in my lab notes? (3)

► Program snippets...

```
chiPhi <- tibble(phi=c(seq(0,0.01, by=0.001),
                        seq(0.01,0.99,by=0.01),
                        seq(0.99,1, by=0.001))) %>%
mutate(chi= 1/(1-2*phi)*log((1-phi)/phi)) %>%
filter(!is.nan(chi))
chiPhib <- chiPhi %>% filter(phi==0.25 | phi==0.75) %>%
  mutate(label=c('$\\phi_b$', '$1-\\phi_b$'))
ggplot(chiPhi) + geom_line(aes(phi, 1/chi)) +
geom_polygon(data=chiPhi %>% add_row(phi=c(0,1), chi=c(Inf,Inf)),
             aes(phi,1/chi, fill='lightgray', alpha=0.5) +
             ylim(0,.7) + xlab("MEP fraction $\\Phi$") + ylab("$1/\\chi$") +
             annotate("text", x=0.5, y=0.6, label="Mixed phase") +
             annotate("text", x=0.5, y=0.2, label="Separated phase") +
             geom_point(data=chiPhib, aes(phi, 1/chi)) +
             geom_line(data=chiPhib, aes(phi, 1/chi, linetype='dashed') +
             geom_text(data=chiPhib, aes(x=phi, y=1/chi, label=label),
                       nudge_y=0.025)
```

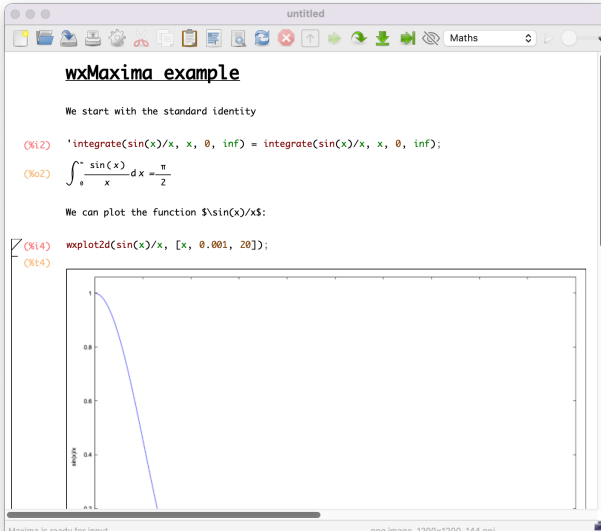
What is in my lab notes? (4)

► ...and their results



Example: Computer algebra notebooks (1)

Many (all) commercial systems have them. Here is a free wxMaxima (<https://wxmaxima-developers.github.io/wxmaxima/>)



The screenshot shows the wxMaxima application window titled "untitled". The window contains a notebook with the following content:

wxmaxima example

We start with the standard identity

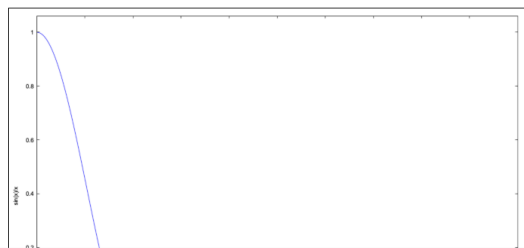
```
(%i2) 'integrate(sin(x)/x, x, 0, inf) = integrate(sin(x)/x, x, 0, inf);
```

(%o2) $\int_0^{\infty} \frac{\sin(x)}{x} dx = \frac{\pi}{2}$

We can plot the function $\sin(x)/x$:

```
(%i4) wxplot2d(sin(x)/x, [x, 0.001, 20]);
```

(%t4)



Maxima is ready for input. png image, 1200x1200, 144 ppi

Example: Computer algebra notebooks (2)

WXMAXIMA EXAMPLE

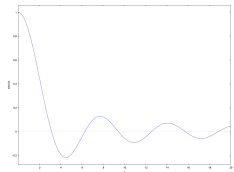
We start with the standard identity

(% i2) `integrate(sin(x)/x, x, 0, inf) = integrate(sin(x)/x, x, 0, inf);`

$$\int_0^{\infty} \frac{\sin(x)}{x} dx = \frac{\pi}{2} \quad (\% o2)$$

We can plot the function $f(x) = \sin(x)/x$:

(% i4) `wplot2d(sin(x)/x, [x, 0.001, 20]);`



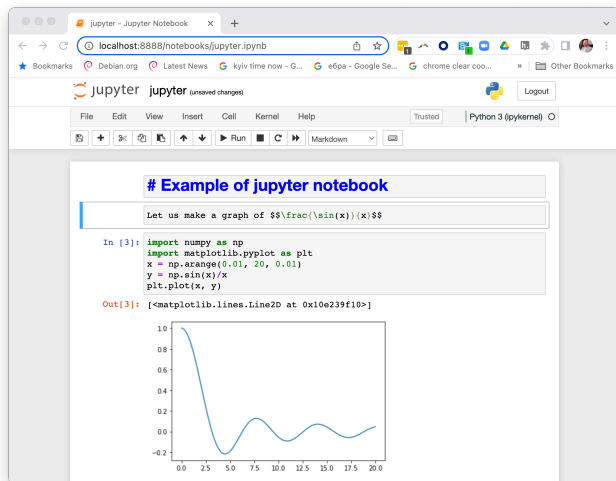
(% o4)

(% o4)

1

Example: Jupyter notebooks (1)

The great Jupyter project (<https://jupyter.org/>)



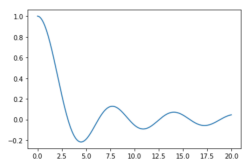
The screenshot shows a web browser window with the URL `localhost:8888/notebooks/jupyter.ipynb`. The Jupyter interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help), a toolbar with icons for undo, redo, and running code, and a status bar showing "Python 3 (ipykernel)".

The notebook content consists of a title cell, a text cell, a code cell, and an output cell.

Example of jupyter notebook

Let us make a graph of $\frac{\sin(x)}{x}$

```
In [3]: import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x, y)
```

Out[3]: [

Example: Jupyter notebooks (2)

jupyter

July 17, 2022

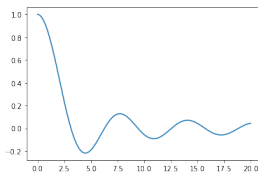
1 Example of jupyter notebook

Let us make a graph of

$$\frac{\sin(x)}{x}$$

```
[3]: import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x, y)
```

[3]: [matplotlib.lines.Line2D at 0x10a239f10]



```
[ ]:
```

1

My (humble) opinions about the examples

wxMaxima: Good for documenting equation manipulations. Not much convenient for everything else.

Jupyter: Good interface, especially when you play with code. Can incorporate many languages other than Python.

But:

- ▶ Only a subset of \LaTeX implemented. No label-ref, bibliography, etc.
- ▶ No support for sketches other than plots.

Common feature: \LaTeX backend. Why not use \LaTeX from the beginning?

My setup

Ideas:

1. I need the features of \LaTeX : bibliographies, numbering, etc.
2. A bunch of `tex` files is easily searched by `grep` and `find`.

A problem: I sometimes play with code and do a lot of plots.

Solution: Use `knitr`.

An aside: \LaTeX and Markdown

An aside: LaTeX and Markdown

Many people use Markdown for

- * notes,
 - * reports,
 - * documents,
 - * some math: $\int_0^{\infty} \sin x/x \, dx = \pi/2$.
-

Markdown: easy to learn, but limited possibilities.

\LaTeX : more difficult to learn, but huge possibilities:
references, bibliographies, sketches, plots...

Preaching to the choir: \LaTeX is a good investment!



Yihui Xie (2015). *Dynamic Documents with R and knitr*. Second edition. Boca Raton; London; New York: Chapman and Hall/CRC. ISBN: 978-1498716963

A great tool for literate programming and literate science (Boris Veytsman (2014). “Book review: Dynamic Documents with R and knitr, by Yihui Xie”. In: *TUGboat* 35.1, pp. 115–119. URL: <http://tug.org/TUGboat/tb35-1/tb109reviews-xie.pdf>).

knitr example (1)

We start from the standard identity

```
\begin{equation}
```

$$\int_0^{\infty} \frac{\sin x}{x} dx = \frac{\pi}{2}$$

```
\end{equation}
```

We also add a simple plot

```
<<device='tikz', fig.width=8, fig.height=3>>=
```

```
data <- tibble(x=seq(0.01, 20, by=0.01)) %>%
```

```
  mutate(y=sin(x)/x)
```

```
ggplot(data) + geom_line(aes(x,y))
```

```
@
```

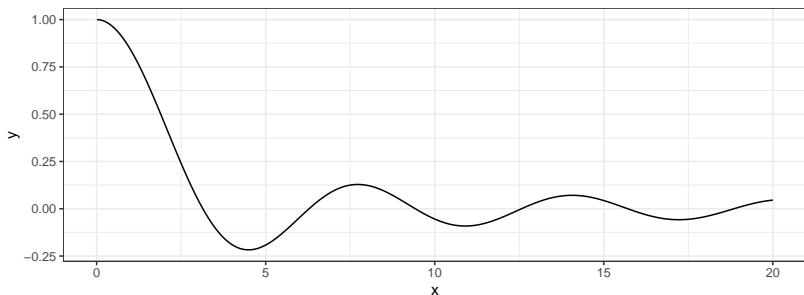
knitr example (2)

We start from the standard identity

$$\int_0^{\infty} \frac{\sin x}{x} dx = \frac{\pi}{2} \quad (1)$$

We also add a simple plot

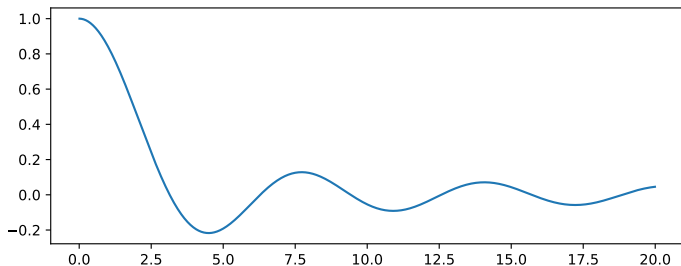
```
data <- tibble(x=seq(0.01, 20, by=0.01)) %>%  
  mutate(y=sin(x)/x)  
ggplot(data) + geom_line(aes(x,y))
```



Not only R!

Here we use `engine='python'` magic

```
import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x,y)
```



Details, tips and tricks

- ▶ Start a project with a directory, `README` and `Makefile` or `Rstudio proj` (or `arara` rules).
- ▶ You may need separate directories for data, etc.
- ▶ Number notes like `001-introduction.rnw`, `002-hypothesis.rnw`, etc.
- ▶ Always use version control!

Examples of my lab notes

1 Introduction

In [1] we have given a number of simple results. The examples we found that the average velocity depends on the geometry of the system. It is interesting to see if we can find a general formula for the average velocity in terms of the geometry of the system. We will consider a system of particles in a box of size L and we will assume that the particles are distributed uniformly in the box.

We start from velocity in a direction.

The velocity in a direction is given by the average of the velocity in that direction. We assume that the particles are distributed uniformly in the box. We will assume that the particles are distributed uniformly in the box.

$$v_x = \frac{1}{L} \int_0^L v_x(x) dx \quad (1)$$

with

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (2)$$

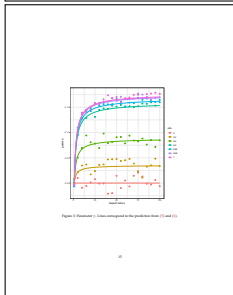
Now we will consider a simple case. We will assume that the particles are distributed uniformly in the box. We will assume that the particles are distributed uniformly in the box.

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (3)$$

We will assume that the particles are distributed uniformly in the box. We will assume that the particles are distributed uniformly in the box.

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (4)$$

and

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (5)$$


2 Simulation

2.1 Simulation function

Let us define a simulation function. We will assume that L is a grid up to the size of the system.

Here is the simulation function:

```

def simulate(L, alpha):
    # ... (code omitted) ...

```

The function `simulate` is a function of L and α with the signature defined by:

```

simulate(L: int, alpha: float) -> float

```

The function `simulate` is a function of L and α with the signature defined by:

```

simulate(L: int, alpha: float) -> float

```

Figure 4: Angle of deflection θ .

3 Introduction

In [1] we have given a number of simple results. The examples we found that the average velocity depends on the geometry of the system. It is interesting to see if we can find a general formula for the average velocity in terms of the geometry of the system. We will consider a system of particles in a box of size L and we will assume that the particles are distributed uniformly in the box.

We start from velocity in a direction.

The velocity in a direction is given by the average of the velocity in that direction. We assume that the particles are distributed uniformly in the box. We will assume that the particles are distributed uniformly in the box.

$$v_x = \frac{1}{L} \int_0^L v_x(x) dx \quad (1)$$

with

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (2)$$

Now we will consider a simple case. We will assume that the particles are distributed uniformly in the box. We will assume that the particles are distributed uniformly in the box.

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (3)$$

We will assume that the particles are distributed uniformly in the box. We will assume that the particles are distributed uniformly in the box.

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (4)$$

and

$$v_x(x) = \frac{1}{L} \int_0^L v_x(x) dx \quad (5)$$

Figure 5: (a) and (b) illustrate the geometry of the system.

Figure 5: (a) and (b) illustrate the geometry of the system.

Figure 5: (a) and (b) illustrate the geometry of the system.

Figure 6: (a) and (b) illustrate the geometry of the system.

Figure 6: (a) and (b) illustrate the geometry of the system.

Figure 6: (a) and (b) illustrate the geometry of the system.

Problems & Solutions

1. Limitations of PDF format: movies & interactive plots are not easy to do! There are solutions, but how reproducible are they? Flash debacle...
2. Speed:
 - ▶ I write prose with the speed I think—*good!*
 - ▶ I program in knitr with the same speed as in IDE—*good!*
 - ▶ I write equations in \TeX slightly slower than with a pen—*ok!*
 - ▶ I write sketches in TikZ (and in PSTricks) much slower than with a pen—*bad!*

Solutions for the sketching speed I am considering:

- ▶ Doodle with a pen, then scan and use `\includegraphics`.
- ▶ Use a program with PDF output.
- ▶ Write TikZ faster.

Final exhortation (standing on the shoulders of a giant)

GO FORTH now and create *beautiful, clear and reproducible laboratory notes!*